**TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN TP.HCM**
**Khoa Hệ Thống Thông Tin**
*Chương trình đào tạo:* Kỹ sư Hệ thống Thông tin

***Bài Thực hành số 6***

*Tên môn học:* ***Dữ liệu lớn***

# I. Tóm tắt bài thực hành

## 1. Yêu cầu lý thuyết

Sinh viên đã được trang bị kiến thức:

- Cấu trúc hệ thống phân tán và framework lập trình Apache Spark
- Đối tượng RDD (Resilient Distributed Dataset) trong Apache Spark
- Lập trình Python với Apache Spark thông qua PySpark
- Các thuật toán máy học

…

## 2. Nội dung

❖ **Ôn tập lại những kiến thức cần thiết**

❖ **Làm quen với thư viện các thuật toán máy học MLLib**

## 3. Kết quả cần đạt

✓ Sử dụng được thư viện Spark MLLib để áp dụng các thuật toán máy học trên dữ liệu lớn.

# II. Ôn tập lại những kiến thức đã học

Sinh viên tham khảo tài liệu ở buổi học trước về việc sử dụng PySpark để lập trình trong Apache Spark, đối tượng RDD.

Sinh viên tham khảo tài liệu ở các môn học trước, tài liệu trên lớp lý thuyết cũng như trên internet để hiểu rõ về các thuật toán máy học.

# III. Yêu cầu bài làm sinh viên

Nội dung thực hành buổi 06 được thực hiện theo từng cá nhân. Sinh viên upload một tập tin **<MSSV>.doc** hoặc **<MSSV>.docx**, nội dung trả lời các bài tập bên dưới.

***Lưu ý: Bài nộp không theo đúng quy định này sẽ không được tính.***

# IV. Làm quen với thư viện các thuật toán máy học (Machine Learning Library – MLLib)

Sinh viên thực thi các đoạn lệnh được đính kèm theo bài thực hành, giải thích ý nghĩa của các câu lệnh và kết quả trả về. Trong đó, đường dẫn đến các tập tin dữ liệu đầu vào tùy thuộc từng máy tính khác nhau, vui lòng điều chỉnh cho phù hợp.

**~ HẾT ~**